

A new jet reconstruction algorithm for lepton colliders

M. Boronat, I. García, M. Vos

IFIC (CSIC/UEVEG), Valencia, Spain

Abstract

We propose a new sequential jet reconstruction algorithm for future lepton colliders at the energy frontier. The Valencia algorithm combines the natural distance criterion for lepton colliders with the greater robustness against backgrounds of algorithms adapted to hadron colliders. Results on a detailed Monte Carlo simulation of $t\bar{t}$ and ZZ production at future linear e^+e^- colliders (ILC and CLIC) with a realistic level of background overlaid, show that it achieves better performance in the presence of background.

Keywords:

jet reconstruction, sequential recombination algorithm, future lepton colliders, ILC, CLIC

1. Introduction

Experiments at lepton and hadron colliders use jet algorithms to cluster the collimated sprays of particles that form in processes with asymptotically free quarks and gluons in the final state. The first modern sequential recombination algorithms were developed for e^+e^- colliders operated at the Z-pole (a detailed historical account is found in Reference [1]). At the heart of the jet algorithm - and crucial to the definition of jets themselves - is a criterion to define the distance between two particles. In popular algorithms used at e^+e^- colliders the distance combines information on the angle between the particles and the energy of (the softest of the two) particles. Sequential recombination algorithms were adapted to the environment at hadron colliders in the early 1990s. At the Large Hadron Collider the large majority of analyses is based on inclusive jet reconstruction with the anti- k_t algorithm [2].

An intense R&D programme exists to develop the technology required for an e^+e^- collider with a center-of-mass energy well beyond that of previous lepton colliders. A linear e^+e^- collider can attain center-of-mass energies from several 100 GeV to several TeV [3, 4]. The possibility of a large circular e^+e^- collider that can reach a center-of-mass

energy of approximately 350 GeV [5] is also explored, as well as a muon collider [6]. Such machines present an environment that differs in several important respects from that encountered at the Z-pole. In this Letter we explore which jet reconstruction algorithms are most suitable for the e^+e^- colliders with a center-of-mass energy from 100 GeV to several TeV.

We start our discussion with a brief recapitulation of the properties of the most popular clustering algorithms in Section 2. We present a proposal for a new jet algorithm in Section 3. In Section 4 the key features of this algorithm are compared to popular algorithms. In Section 5 the Monte Carlo simulation setup that we used to benchmark the performance of the algorithms is introduced. Finally, in Sections 6 and 7 we present the results for top quark pair and di-boson (ZZ) production at the ILC and CLIC, in a realistic environment including the relevant background. In Section 8 we summarize the most important findings of this work.

2. Overview of jet reconstruction algorithms based on sequential recombination

The first modern clustering algorithm with a simple sequential recombination scheme algo-

rithm is the JADE algorithm developed in the middle of the 1980s [7, 8]. The distance y_{ij} assigned to any pair of particles i and j is given by:

$$y_{ij} = \frac{E_i^2 E_j^2}{Q^2} (1 - \cos \theta_{ij}) \quad (1)$$

where E_i and E_j denote the energy of the two particles, Q is the total energy of the event, and θ_{ij} is the angle between the two particles. At each step the algorithm merges the pair of particles with the smallest distance y_{ij} . This process continues until the smallest distance exceeds a value y_{cut} (*inclusive* clustering) or a previously defined number of jets is obtained (*exclusive* clustering).

In the Durham or $e^+e^- k_t$ algorithm [9] used extensively at LEP and SLC the distance between particles i and j is modified to depend on the minimum of the energies E_i and E_j , rather than the product $E_i E_j$:

$$d_{ij} = 2 \min(E_i^2, E_j^2) (1 - \cos \theta_{ij}) \quad (2)$$

For sufficiently small angles the distance reduces to the transverse momentum squared of the softer particle relative to the harder one. The distance measure is thus proportional to the squared inverse of the splitting probability for one parton k into partons i and j in the soft and collinear limit.

Jet reconstruction at hadron colliders presents a number of additional difficulties. The incoming beams radiate gluons that can form jets. Only a fraction of the energy of the composite projectiles is transferred in the hard parton-parton process and a hadron remnant continues to travel down the beam pipe. An important consequence is that the system formed by the reaction products is typically not at rest in the laboratory frame¹. Clustering algorithms were adapted to meet these challenges in the 1990s.

¹ For di-jet production at the LHC $\beta_z = v_z/c$ of the di-jet system is very close to 1 and even a massive system such as a top quark pair acquires a typical $\beta_z = 0.5$. In contrast, for processes such as $e^+e^- \rightarrow ZH(\gamma)$ (Higgsstrahlung) at $\sqrt{s} = 250$ GeV and $e^+e^- \rightarrow t\bar{t}(\gamma)$ at 500 GeV β_z is smaller than 0.1 in 95% and 90% of the events, respectively. The exception to the rule is the $2 \rightarrow 2$ process $e^+e^- \rightarrow f\bar{f}(\gamma)$, with f any fermion lighter than the Z-boson, where ISR (return-to-the-Z) plays an important role.

The first important modification of the algorithms is the addition of so-called *beam jets*, introduced in Reference [10]. Any particle with a beam distance $d_{iB} = p_{Ti}^{2n}$ smaller than any d_{ij} is not merged with any other particle, but is associated to the beam jet. These are not considered part of the visible final state. Thus, the soft, collinear radiation emitted by the incoming hadrons and the hadron remnant travelling in the very forward and backward direction are discarded.

To cope with the boost along the beam direction, analyses at hadron colliders replace the particle energy E_i with its transverse momentum p_{Ti} and the angular distance between the particles $(1 - \cos \theta_{ij})$ with $\Delta R_{ij} = \sqrt{(\Delta\phi)^2 + (\Delta y)^2}$, where y denotes the rapidity. In the longitudinally invariant k_t algorithm [11, 12] the distance criterion is based on the same observables “to improve the factorization properties [of the algorithm] and [achieve] closer correspondence to experimental practice [...]” [11]. We rewrite the generic inter-particle distance as follows:

$$d_{ij} = \min(p_{Ti}^{2n}, p_{Tj}^{2n}) \frac{\Delta R_{ij}^2}{R^2} \quad (3)$$

where R is the radius parameter. Setting n in the exponent to 1 yields the longitudinally invariant k_t algorithm. Alternative choices of the exponent yield the Cambridge-Aachen algorithm ($n=0$), or the anti- k_t algorithm ($n=-1$), the default jet reconstruction algorithm at the LHC.

Finally, one can add beam beam jets to the k_t algorithm for e^+e^- experiments. This yields an algorithm we refer to as the generic $e^+e^- k_t$ algorithm, with inter-particle distance:

$$d_{ij} = \min(E_i^2, E_j^2) (1 - \cos \theta_{ij}) / (1 - \cos R) \quad (4)$$

and beam distance given by $d_{iB} = E_i^2$.

3. The Valencia jet algorithm

Background levels at hadron colliders form an important consideration in the design of jet algorithms. The *pile-up* of several tens of minimum bias events on each bunch crossing at the LHC is a serious challenge that has led to a large body of work on mitigation and correction

methods. In comparison, previous lepton colliders, such as LEP or SLD, presented an environment with essentially negligible background. Future lepton colliders are in between these two extremes. While very far from the background levels of the LHC, detailed studies of the $\gamma\gamma \rightarrow \text{hadrons}$ background at the ILC or CLIC have shown a non-negligible impact on the jet reconstruction performance [4, 13]. Among several proposals to mitigate its effect, the use of the longitudinally invariant k_t algorithm, intended for hadron colliders, has led to the greatest improvement of the robustness.

We propose a new clustering jet reconstruction algorithm for future e^+e^- colliders that maintains a Durham-like distance criterion based on [energy, polar angle] (as opposed to [transverse momentum, rapidity] in the hadron collider algorithm) and can compete with the robustness against background of the longitudinally invariant k_t algorithm. The algorithm has the following inter-particle distance:

$$d_{ij} = \min(E_i^{2\beta}, E_j^{2\beta})(1 - \cos \theta_{ij})/R^2 \quad (5)$$

For $\beta = 1$ the distance is given by the transverse momentum squared of the softer of the two particles relative to the harder one, as in the Durham algorithm. Note that we have redefined the meaning of the radius parameter R with respect to the generalized e^+e^- algorithm with beam jets. The R^2 in the numerator yields greater freedom than the $1 - \cos R$, that is limited to the interval $[0, 2]$.

The beam distance of the Valencia algorithm is:

$$d_{iB} = p_T^{2\beta} \quad (6)$$

For $\beta = 1$ this combination of inter-particle and beam distance metrics is similar to that of the k_\perp algorithm proposed in Ref. [10], with the difference that $d_{iB} = p_{Ti}^2 = E_i^2 \sin^2 \theta_{iB}$, whereas in Ref. [10] it was given by $2E_i^2(1 - \cos \theta_{iB})$.

The Valencia algorithm is available as a plug-in for the FastJet [14, 15] package. The code can be obtained from the “contrib” area [16].

4. Comparison of the distance criteria of sequential recombination algorithms

The choice of distance criterion defines the essence of the jet algorithm and has profound im-

plications on its performance in a given environment. The differences between the various algorithms are most easily visualized as follows. We calculate the distance between two test particles with an energy of 1 GeV emitted at a fixed relative angle of 100 mrad. The leftmost plot in Figure 1 shows how the distance between the two particles evolves as the system is scanned from the central detector ($\cos \theta = 0$) to the forward region ($\cos \theta = 1$).

The distance d_{ij} of the generic $e^+e^- k_t$ algorithm is independent of polar angle, as shown in Figure 1. The same holds for the Valencia algorithm proposed here, but generally not for algorithms used at hadron colliders. Two effects come into play. For two particles separated by a given polar angle, the pseudo-rapidity difference $\Delta\eta$ grows larger in the forward region. At the same time the distance between two particles with energy E decreases as p_T is reduced. The net effect for the k_t algorithm is a sharp decrease of the distance in the forward region.

The relation between the inter-particle distance d_{ij} and the beam distance d_{iB} governs the relative *attraction* of beam jets and final-state jets and is therefore a crucial property for the performance in environments with significant background. The ratio $\frac{d_{ij}}{d_{iB}}$ is shown as a function of polar angle in the central plot in Figure 1. As might be expected from the functional form in Equation 4, the ratio is flat for e^+e^- algorithms (Durham). For the longitudinally invariant k_t algorithm, on the other hand, the ratio rises steeply in the forward region. For the Valencia algorithm with $\beta = 1$ we obtain very similar behaviour to longitudinally invariant k_t .

The steep rise in $\frac{d_{ij}}{d_{iB}}$ at $\cos \theta \sim 1$ penalizes relatively isolated particles in the forward and backward directions, that are likely due to background processes. The exponent β introduced in the Valencia algorithm gives a handle to enhance or diminish the increase of the $\frac{d_{ij}}{d_{iB}}$ ratio in the forward region, as shown in Figure 1. Thus, we have a handle to *tune* the background rejection that is independent of the parameter R that governs the jet radius.

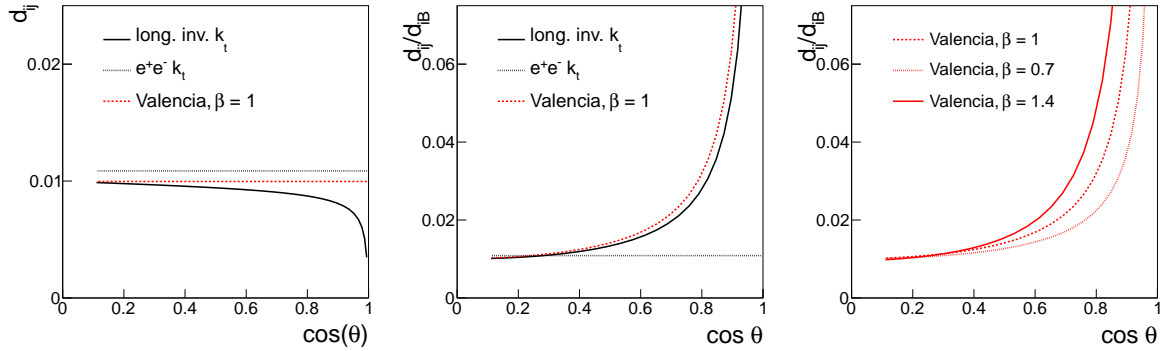


Figure 1: The dependence of the inter-particle distance d_{ij} of two test particles emitted at fixed angular distance and the ratio of d_{ij} to the beam distance d_{iB} with the polar angle θ . Results are presented for several clustering jet reconstruction algorithms discussed in the text.

5. Monte Carlo simulation

The performance of the different algorithms is compared for $t\bar{t}$ and ZZ production at a linear e^+e^- collider with $\sqrt{s} = 500$ GeV. Samples are generated with WHIZARD [17]. The response of the ILD detector [18] is simulated with GEANT4 [19].

The background considered in this study is due to multi-peripheral $\gamma\gamma \rightarrow \text{hadrons}$ production². The background events are overlaid on the signal using a mechanism similar to that used for pile-up at the LHC. For a 500 GeV e^+e^- collider less than one $\gamma\gamma \rightarrow \text{hadrons}$ events is produced per bunch crossing.

The impact of the background on the output of the detector is quite different at CLIC and the ILC. At CLIC bunches are spaced by 500 picoseconds and detector systems are expected to integrate the background of a number of subsequent bunch crossings. In this study the background corresponding to a large number of bunch crossings is overlaid (300 for 500 GeV operation, 60 for 3 TeV). The much larger bunch spacing at the ILC allows the detector to distinguish single bunch crossings, such that less than one $\gamma\gamma \rightarrow \text{hadrons}$ event is overlaid (on average) on each signal event.

In the event reconstruction the information of the tracking system and the calorimeters is combined to form particle-flow objects with the Pan-

dora [20] algorithm. In the CLIC studies particle flow objects are selected using a set of timing cuts, corresponding to the nominal selection of Ref. [13].

6. Top quark pair production at a 500 GeV ILC

We study the performance of several jet algorithms in the study of $t\bar{t}$ production at the ILC of Ref. [21]. The Monte Carlo sample includes all six-fermion processes that produce a “lepton + jets” final state: $e^+e^- \rightarrow b\bar{b}l^\pm\nu_l q\bar{q}'$.

Reconstruction of the event involves charged lepton reconstruction and removal of the corresponding energy, the reconstruction of exactly four jets (exclusive jet clustering with $N = 4$) and flavour tagging, described in detail in Ref. [21]: The two jets with poorest score in the b-tagging algorithm are combined to form the W -boson candidate. The hadronic top candidate is constructed by adding the remaining (b-)jet that minimizes a χ^2 based on the hadronic top quark candidate mass and energy, the b-jet energy in the top quark rest frame and the angle between W -boson and b-quark.

We consider four jet reconstruction algorithms: the Durham algorithm, the generic $e^+e^- k_t$ algorithm with beam jets with $R = 1$, the longitudinally invariant k_t algorithm with $R = 1.5$ and the Valencia algorithm with $R = 1.2$ and $\beta = 0.8$. The choice of parameters corresponds to the optimal setting determined in a scan over a broad range of parameters. The resolution of the measurements of the

²A further source of background, pair production from beamstrahlung photons is ignored in this discussion.

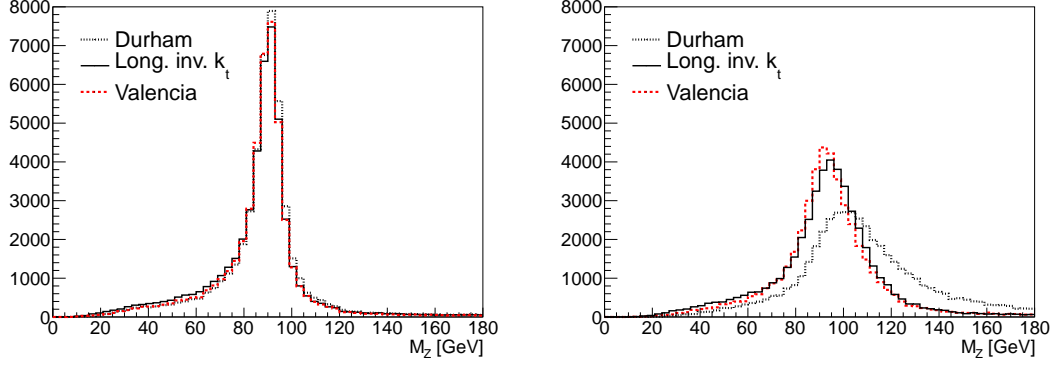


Figure 2: The reconstructed Z-boson mass distribution for $ZZ \rightarrow q\bar{q}q'\bar{q}'$ events at a 500 GeV CLIC. No backgrounds are added in the leftmost plot. The results on the rightmost plot correspond to the same events with the $\gamma\gamma \rightarrow \text{hadrons}$ background corresponding to 300 bunch crossings overlaid on the signal, where each bunch crossing contains approximately 0.3 $\gamma\gamma \rightarrow \text{hadrons}$ events.

RMS ₉₀ [GeV]	E_{4j}	E_W	m_W	E_t	m_t
Durham	23.2	19.6	20.3	19.5	21.4
$e^+e^- k_t$	25.6	20.8	21.6	20.5	22.8
long. inv. k_t	21.7	18.4	18.9	18.4	20.1
Valencia	21.4	18.0	18.8	18.2	20.0

Table 1: The Root Mean Square of the central 90% of the events (RMS90) for five observables reconstructed in $t\bar{t}$ events at a 500 GeV ILC: the energy of the system formed by the four jets, the energy and mass of the hadronic W-boson and the energy and mass of the hadronic top quark.

energy of the four jets, of the energy and mass of the hadronic W-boson and hadronic top quark candidate are given in Table 1.

The results show a clear advantage of the algorithms with a d_{ij}/d_{iB} ratio that increases in the forward and backward region of the experiment. Even with the rather modest background level at the ILC the longitudinally invariant k_t algorithm and the algorithm proposed in this Letter achieve a 10-15% better resolution and a smaller bias than the e^+e^- algorithms.

7. Di-boson production at CLIC

The $e^+e^- \rightarrow ZZ$ process is studied in the CLIC environment to enable comparison with the first detailed studies of the impact of background on jet reconstruction at future lepton colliders in Ref. [13] and the CLIC CDR [4].

We select $e^+e^- \rightarrow ZZ \rightarrow q\bar{q}q'\bar{q}'$ events. Events with Z-bosons emitted in the very forward direction (with polar angle $|\cos\theta| > 0.99$), where the beam pipe may have a profound impact are discarded, as well as events where the Z-bosons are very far from their mass shell ($|m(q\bar{q}) - m_Z| > 30$ GeV).

Exactly four jets are reconstructed and the di-jet combinations are selected that minimize the following χ^2 :

$$\chi^2 = \frac{(E_{Z1} - E_{Z2})^2}{(250 \text{ GeV})^2} + \frac{(m_{Z1} - m_{Z2})^2}{(91 \text{ GeV})^2} + \frac{\angle(Z_1, Z_2)}{(\pi)^2}.$$

The Z boson candidate mass distribution is shown in Figure 2. Numerical results are given in Table 2.

$\sqrt{s} = 500$ GeV, no background overlay			
[GeV]	m_Z	σ_Z	RMS ₉₀
Durham	90.6	5.4	13.8
long. inv. k_t	90.4	5.3	14.3
Valencia	90.3	5.2	12.5
$\sqrt{s} = 500$ GeV, 0.3 $\gamma\gamma \rightarrow \text{hadrons}$ events/BX			
[GeV]	m_Z	σ_Z	RMS ₉₀
Durham	101.1	13.6	28.8
long. inv. k_t	95.1	10.9	17.9
Valencia	93.1	10.2	17.1

Table 2: The center and width - from a Gaussian fit - of the reconstructed Z-boson mass peak in ZZ events at a 500 GeV CLIC. The third column lists the RMS90 estimate.

In the background-free case all three algorithms achieve a narrow Z-boson mass peak. The impact of the overlaid background is rather pronounced for the Durham algorithm. The peak position shifts by approximately 10 GeV and broadens considerably. Both the longitudinally invariant k_t algorithm and the Valencia algorithm show considerably better performance under these conditions.

8. Conclusions

We propose a jet algorithm that offers robust performance in the presence of the mild background levels expected at lepton colliders, while retaining the natural inter-particle distance criterion in the [energy, angle] basis (as opposed to the [transverse momentum, rapidity] basis of hadron collider algorithms). The algorithm is further generalised with a variable exponent that allows to tune the background rejection for the specific requirements of a given analysis. We have benchmarked the performance of several algorithms in a full Monte Carlo simulation studies of $t\bar{t}$ and ZZ production at the ILC and CLIC. We find that the Valencia algorithm performs better than the sequential recombination algorithms used at previous lepton colliders.

Acknowledgement

The authors would like to thank Gavin Salam for helpful suggestions and guidance creating the plugin code and Bryan Webber for his careful reading of the manuscript.

References

- [1] S. Moretti, L. Lonnblad and T. Sjostrand, *New and old jet clustering algorithms for electron - positron events*, *JHEP* **9808** (1998) 001 [hep-ph/9804296].
- [2] M. Cacciari, G. P. Salam and G. Soyez, *The Anti- $k(t)$ jet clustering algorithm*, *JHEP* **0804** (2008) 063 [0802.1189].
- [3] H. Baer, T. Barklow, K. Fujii, Y. Gao, A. Hoang *et al.*, *The International Linear Collider Technical Design Report - Volume 2: Physics*, 1306.6352.
- [4] L. Linssen, A. Miyamoto, M. Stanitzki and H. Weerts, *Physics and Detectors at CLIC: CLIC Conceptual Design Report*, 1202.5940.
- [5] M. Bicer, H. Duran Yildiz, I. Yildiz, G. Coignet, M. Delmastro *et al.*, *First Look at the Physics Case of TLEP*, 1308.6176.
- [6] Y. Alexahin, C. M. Ankenbrandt, D. B. Cline, A. Conway, M. A. Cummings *et al.*, *Muon Collider Higgs Factory for Snowmass 2013*, 1308.2143.
- [7] **JADE Collaboration** Collaboration, W. Bartel *et al.*, *Experimental Studies on Multi-Jet Production in e^+e^- Annihilation at PETRA Energies*, *Z.Phys.* **C33** (1986) 23.
- [8] **JADE Collaboration** Collaboration, S. Bethke *et al.*, *Experimental Investigation of the Energy Dependence of the Strong Coupling Strength*, *Phys.Lett.* **B213** (1988) 235.
- [9] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock and B. Webber, *New clustering algorithm for multi - jet cross-sections in e^+e^- annihilation*, *Phys.Lett.* **B269** (1991) 432–438.
- [10] S. Catani, Y. L. Dokshitzer and B. Webber, *The K^- perpendicular clustering algorithm for jets in deep inelastic scattering and hadron collisions*, *Phys.Lett.* **B285** (1992) 291–299.
- [11] S. Catani, Y. L. Dokshitzer, M. Seymour and B. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, *Nucl.Phys.* **B406** (1993) 187–224.
- [12] S. D. Ellis and D. E. Soper, *Successive combination jet algorithm for hadron collisions*, *Phys.Rev.* **D48** (1993) 3160–3166 [hep-ph/9305266].
- [13] J. Marshall, A. Muennich and M. Thomson, *Performance of Particle Flow Calorimetry at CLIC*, *Nucl.Instrum.Meth.* **A700** (2013) 153–162 [1209.4039].
- [14] M. Cacciari and G. P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, *Phys.Lett.* **B641** (2006) 57–61 [hep-ph/0512210].
- [15] M. Cacciari, G. P. Salam and G. Soyez, *FastJet User Manual*, *Eur.Phys.J.* **C72** (2012) 1896 [1111.6097].
- [16] “ValenciaJetAlgorithm plug-in for fastjet.” <https://fastjet.hepforge.org/contrib/>.
- [17] W. Kilian, T. Ohl and J. Reuter, *WHIZARD: Simulating Multi-Particle Processes at LHC and ILC*, *Eur.Phys.J.* **C71** (2011) 1742 [0708.4233].
- [18] T. Behnke, J. E. Brau, P. N. Burrows, J. Fuster, M. Peskin *et al.*, *The International Linear Collider Technical Design Report - Volume 4: Detectors*, 1306.6329.
- [19] **GEANT4** Collaboration, S. Agostinelli *et al.*, *GEANT4: A Simulation toolkit*, *Nucl.Instrum.Meth.* **A506** (2003) 250–303.
- [20] J. Marshall and M. Thomson, *The Pandora software development kit for particle flow calorimetry*, *J.Phys.Conf.Ser.* **396** (2012) 022034.
- [21] M. Amjad, M. Boronat, T. Frisson, I. Garcia, R. Poschl *et al.*, *A precise determination of top quark electro-weak couplings at the ILC operating at $\sqrt{s} = 500$ GeV*, 1307.8102.